JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11)Publication number: 07274965
(43)Date of publication of application: 24.10.1995

(51)Int.Cl.
C12N 15/00
// G01N 33/50
G06F 12/00

(21)Application number: 06275336
(22)Date of filing: 09.11.1994

(71)Applicant: KOKURITSU IDENGAKU KENKYUSHO
FUJITSU LTD

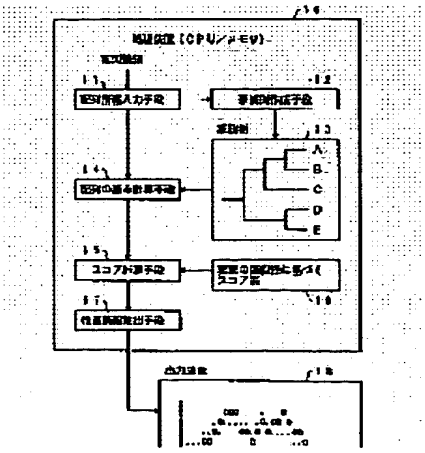
(72)Inventor: GOJIYOUBORI TAKASHI
TATENO YOSHIO
IKEO KAZUO
KAWANISHI YUICHI
KAWAI MASATO

(30)Priority
Priority number: 05283329 Priority date: 12.11.1993 Priority country: JP

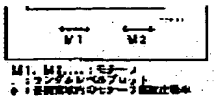
(54) APPARATUS FOR EXTRACTION TREATMENT OF MOTIF OF GENE AND TREATING METHOD

(57)Abstract:

PURPOSE: To provide an apparatus for the extraction treatment of the motif of a gene, capable of mechanically (automatically) extracting a regularity characteristic to a sequence and enabling the specification of a gene function from gene sequence information and to provide a method for the extraction treatment.
CONSTITUTION: Plural gene sequences are inputted, each sequence is weighted from alignment data on the sequences based on a phylogenetic tree, the score of each site is calculated from the weight of each sequence and the analogy of amino acids and a part having large score is automatically



extracted as a motif having a regularity characteristic to the sequence.



LEGAL STATUS

[Date of request for examination] 08.08.1997

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（A）　(11)特許出願公開番号

特開平7－274965

(43)公開日　平成７年(1995)10月24日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｃ１２Ｎ　15/00				
// Ｇ０１Ｎ　33/50	Ｐ			
Ｇ０６Ｆ　12/00	５１０　Ｂ	7608－5B		
		9281－4B	Ｃ１２Ｎ　15/ 00　Ｚ	

審査請求　未請求　請求項の数９　ＯＬ　（全 19 頁）

(21)出願番号　特願平6－275336

(22)出願日　平成６年(1994)11月９日

(31)優先権主張番号　特願平5－283329

(32)優先日　平５(1993)11月12日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　593206872
国立遺伝学研究所長
静岡県三島市谷田1111番地

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中1015番地

(72)発明者　五條堀　孝
静岡県三島市谷田1111番地　国立遺伝学研究所内

(72)発明者　舘野　義男
静岡県三島市谷田1111番地　国立遺伝学研究所内

(74)代理人　弁理士　伊東　忠彦

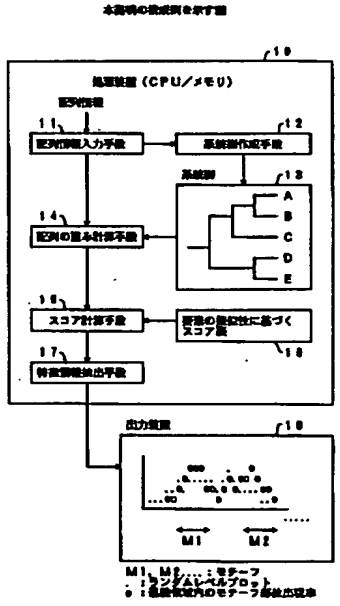
最終頁に続く

(54)【発明の名称】　遺伝子のモチーフ抽出処理装置及び処理方法

(57)【要約】

【目的】　遺伝子配列情報から遺伝子機能を特定する、配列に特徴的な規則性を抽出する遺伝子のモチーフ抽出処理装置および処理方法に関し、遺伝子配列情報をもとに機械的に（自動的に）モチーフを抽出することを目的とする。

【構成】　入力した複数の遺伝子配列のアライメントデータから、進化系統樹に基づく各配列への重み付けを行い、各配列の重みとアミノ酸の類似性とから各部位のスコアを計算し、スコアが大きい部分を配列に特徴的な規則性であるモチーフとして自動抽出する。



1

【特許請求の範囲】

【請求項１】　遺伝子配列情報から遺伝子機能を特定する特徴的な規則性を抽出する処理装置であって、
複数の遺伝子配列に関する進化系統樹（１３）の枝の長さから各配列の重みを計算する配列の重み計算手段（１４）と、
配列の各部位毎に、その部位において出現する配列要素の類似性の度合を示すスコアを前記重みを用いて計算するスコア計算手段（１５）と、
計算されたスコアに基づいて遺伝子配列における特徴的な規則性を有する部分をモチーフとして抽出する特徴情報抽出手段（１７）とを備えた、
遺伝子のモチーフ抽出処理装置。

【請求項２】　請求項１記載の遺伝子のモチーフ抽出処理装置において、
複数の遺伝子配列のアライメントデータをもとに遺伝子配列間の相違度に基づく進化系統樹（１３）を作成する系統樹作成手段（１２）を更に備えた、
遺伝子のモチーフ抽出処理装置。

【請求項３】　請求項１又は２記載の遺伝子のモチーフ抽出処理装置において、
前記スコア計算手段（１５）は、前記配列の重み計算手段（１４）によって計算された各配列の重みと、予め配列要素の種類に応じて求められている配列要素間の類似性情報（１６）とに基づいてスコアを計算する手段を有する、
遺伝子のモチーフ抽出処理装置。

【請求項４】　請求項１～３のうちいずれか一項記載の遺伝子のモチーフ抽出処理装置において、
前記特徴情報抽出手段（１７）は、前記スコア計算手段（１５）によって計算されたスコアの値が所定の閾値または設定された閾値を超えた場合に、その部位をモチーフ部位として抽出し出力する手段を有する、
遺伝子のモチーフ抽出処理装置。

【請求項５】　請求項４記載の遺伝子のモチーフ抽出処理装置において、
前記特徴情報抽出手段（１７）は、所定の連続領域幅又は設定した連続領域幅でモチーフ部位の出現率を計算し、その値が所定のランダムレベル又は設定したランダムレベルを超えた場合に、その連続領域をモチーフ領域として抽出すると共に、隣合う領域がともにモチーフ領域の場合は、それらの領域を１つのモチーフ領域として出力する手段を更に有する、
遺伝子のモチーフ抽出処理装置。

【請求項６】　請求項５記載の遺伝子のモチーフ抽出処理装置において、
少なくともモチーフ部位の出現率やランダムレベルの値をプロットしてグラフ表示することにより前記特徴情報抽出手段（１７）からの特徴情報を出力する出力手段（１８）を更に有する、

2

遺伝子のモチーフ抽出処理装置。

【請求項７】　計算機により遺伝子配列情報から遺伝子機能を特定する特徴的な規則性を抽出する処理方法であって、
抽出対象となる複数の遺伝子配列のアライメントデータを入力する処理過程と、
それをもとに進化系統樹を作成する処理過程と、
その系統樹における枝の長さから各配列の重みを計算する処理過程と、
計算された各配列の重みと、予め配列要素の種類の応じて求められている配列要素間の類似性情報とに基づいて、各部位のスコアを計算する処理過程と、
計算されたスコアの値が所定の閾値または設定された閾値を超えた場合に、その部位をモチーフ部位として抽出する処理過程とを有する、
遺伝子のモチーフ抽出処理方法。

【請求項８】　請求項７記載の遺伝子のモチーフ抽出処理方法において、
所定の連続領域幅又は設定した連続領域幅でモチーフ部位の出現率を計算し、その値が所定のランダムレベル又は設定したランダムレベルを超えた場合に、その連続領域をモチーフ領域として抽出すると共に、隣合う領域がともにモチーフ領域の場合は、それらの領域を１つのモチーフ領域として出力する処理過程を更に有する、
遺伝子のモチーフ抽出処理方法。

【請求項９】　請求項８記載の遺伝子のモチーフ抽出方法において、
少なくともモチーフ部位の出現率やランダムレベルの値をプロットしてグラフ表示する前記領域をモチーフとして抽出する処理過程からの特徴情報を出力する処理過程を更に有する、
遺伝子のモチーフ抽出処理方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は遺伝子のモチーフ抽出処理装置及び処理方法に係り、特に与えられた複数の遺伝子配列情報の比較からそれらの配列間の保存部位であるモチーフを抽出する遺伝子のモチーフ抽出処理装置及び処理方法に関する。近年の遺伝子工学の進歩に伴い、ＤＮＡ配列やアミノ酸配列で表現される遺伝子配列情報データベースが急増している。また、ヒトゲノム計画などのように、特定の生物の遺伝子配列を全て解明しようという試みが世界的規模で行われており、遺伝子配列情報は今後も急激に増加することが予想される。

【０００２】これらの遺伝子配列の中には、配列情報は明らかになっているが、その機能や構造に関しては未知であるものも多い。このような遺伝子の機能や構造を、その配列情報から予測するために有効な方法として、配列に特徴的な規則性であるモチーフの検索がある。そのために、配列が既知のものから多くのモチーフを抽出す

る技術が必要とされる。

【０００３】

【従来の技術】従来、遺伝子配列において遺伝子機能を特定する、配列に特徴的な規則性を示すモチーフは、実験や文献での報告に基づいて決定されてきた。このようなモチーフを登録したデータベースとして、ＰＲＯＳＩＴＥが知られている。ところで、一般に、遺伝子配列の中で、機能的に重要な部位（サイト）は変わりにくいことが知られている。このことを利用すれば、複数の遺伝子配列の比較から、保存領域としてモチーフを抽出することができる。しかしながら、従来、遺伝子配列の比較からモチーフを抽出する手法は確立されていない。

【０００４】

【発明が解決しようとする課題】実験等により人為的にモチーフを決定するのは、大変な作業である。そこで、遺伝子配列の比較からモチーフを機械的に抽出することができれば、遺伝子機能の解明等に有効な多くの情報を得ることができると考えられる。しかし、単に複数の遺伝子配列の各部位を比較し、各部位の類似性を調べていく手法を採った場合、次のような問題がある。

【０００５】つまり、抽出対象とする複数の遺伝子配列情報が特定の種類の生物に偏った場合、抽出しようとする規則性に偏りが生じる。例えば、人間の遺伝子配列情報、猿の遺伝子配列情報、馬の遺伝子配列情報、・・・等の高等生物の遺伝子配列情報が多数あり、それより下等な生物の遺伝子配列情報が少ない配列情報群について、各部位の類似性からモチーフを抽出しようとした場合、類似性の高い部分が進化においてあまり変化していない保存領域であるとは必ずしも認定することはできず、モチーフとして抽出する保存領域の認定が誤りが生じる可能性がある。この逆の場合も同様である。

【０００６】本発明は上記問題点の解決を図り、複数の遺伝子配列情報をもとに、機械的に（自動的に）モチーフを抽出することを目的とする。

【０００７】

【課題を解決するための手段】図１は本発明の構成例を示す図である。図１において、１０はＣＰＵ及びメモリ等からなる処理装置である。配列情報入力手段１１は、モチーフの抽出対象となる複数の遺伝子配列のアライメントデータを入力する手段である。系統樹作成手段１２は、配列情報入力手段１１によって入力した複数の遺伝子配列のアライメントデータをもとに遺伝子配列間の相違度に基づく進化系統樹１３を作成する手段である。なお、系統樹１３は、例えば古生物学的な情報等を用いて予め作成しておくようにしてもよい。

【０００８】配列の重み計算手段１４は、系統樹１３の枝の長さから各配列の重みを計算する手段である。スコア計算手段１５は、配列の各部位毎に、その部位において出現する配列要素の類似性の度合を示すスコアを、配列の重みと、予め配列要素の種類に応じて求められている要素の類似性に基づくスコア表１６とに基づいて計算する手段である。

【０００９】特徴情報抽出手段１７は、計算されたスコアに基づいて遺伝子配列における特徴的な規則性を有する部分をモチーフとして抽出し、ディスプレイやプリンタ等の出力装置１８に出力させる手段である。特に、特徴情報抽出手段１７は、スコア計算手段１５によって計算されたスコアの値が所定の閾値または設定された閾値を超えた場合に、その部位をモチーフ部位として抽出する。又、特徴情報抽出手段１７は、所定の連続領域幅又は設定した連続領域幅でモチーフ部位の出現率を計算し、その値が所定のランダムレベル又は設定したランダムレベルを超えた場合には、その連続領域をモチーフ領域とし、隣合うモチーフ領域を１つのモチーフ領域とする。これらの計算結果等が出力装置１８へ出力される。出力装置１８は、モチーフの出現率やランダムレベルの値等をプロットしてグラフ表示を出力する。

【００１０】

【作用】本発明の遺伝子のモチーフ抽出処理装置は、マルティプルアライメントデータを入力データとし、各部位においてアライメントデータを構成する配列中で高度に保存されているアミノ酸をモチーフとして出力する。ただし、進化的に近縁な配列が存在することによる、アミノ酸の出現頻度の偏りを補正するために、アライメントデータに基づき系統樹１３を作成し、系統樹１３の枝長や形から、各遺伝子配列に対する重み付けを行う。更に、性質の似たアミノ酸の出現を許容するために、アミノ酸の類似性に基づいて計算されたスコア表１６を用いて、各部位でのスコアを計算する。ここで求められたスコアが高いほど、その部位では、アミノ酸が高度に保存されていることを示す。

【００１１】更に、モチーフ部位を抽出するために、スコアの閾値を設定する操作を行う。又、ここで設定した閾値を超えるスコアを示した部位をモチーフ部位として抽出する。そして、モチーフ領域を限定するために、領域幅とランダムレベルとを設定する操作を行う。ここで設定した領域幅内でのモチーフ部位の出現率がランダムレベルを超える値を示した場合、その領域をモチーフ領域とみなす。又、隣合うモチーフ領域は、１つのモチーフ領域とみなす。

【００１２】

【実施例】以下、図面を参照しつつ、本発明の実施例をアミノ酸配列で表される遺伝子配列情報を例にして説明する。図２は本発明の実施例の処理フローチャートである。以下の説明における処理（ａ）～（ｋ）は、図２に示す処理（ａ）～（ｋ）に対応する。

【００１３】図３に示す５本の遺伝子配列Ａ～Ｅからなるアライメントデータを考える。アルファベット一文字がひとつのアミノ酸に対応し、配列長の＊はギャップを表す。

（ａ）アライメントデータ入力
配列情報入力手段１１は、図３に示す配列Ａ～Ｅのアライメントデータを入力する。配列情報がよく似た配列が多数存在する場合、各部位を代表するアミノ酸をその出現頻度から求めると偏りが生ずる。そこで、以下の処理では、入力したアライメントデータから系統樹１３を作成し、系統樹１３の枝長や形をもとに、各遺伝子配列に対する重み付けの計算を行う。その計算結果を用い、各配列に対して重み付けを行うことで、偏りを補正する。

【００１４】（ｂ）系統樹作成
系統樹作成手段１２による系統樹１３の作成には、例えばＵＰＧ（UnweightedPair-Group Clustering）法を用いる。他の作成方法を用いてもよい。本実施例では、具体的には系統樹の作成を以下のように行う。先ず、アライメントデータをもとに、遺伝子配列間の相違度を求める。相違度は２本ずつの配列を組にして、それら配列間のアミノ酸の置換数として計算される。計算式はアミノ酸置換数を求める時に一般的に使われる次の式（１）を用いる。

【００１５】

$$K=-\log(1-p) \quad \cdots \text{式}(1)$$

ここで、Ｋはアミノ酸数置換数、ｐは２本の配列間で異なるアミノ酸を持つ部位の割合である。また、ギャップを含む部位については計算から除外する。式（１）により、全ての２本の配列の組、即ち（Ａ，Ｂ），（Ａ，Ｃ），・・・（Ａ，Ｅ），（Ｂ，Ｃ），・・・，（Ｃ，Ｄ），（Ｃ，Ｅ），（Ｄ，Ｅ）の組について相違度を計算する。また、この相違度をV_{AB}，V_{AC}，・・・，V_{DE}と表すと、相違度V_{AB}，V_{AC}，・・・，V_{DE}の中で最小のものを選び、その組を結び付ける。この例では、配列Ｄと配列Ｅが結び付けられる。この相違度を枝の長さとする。

【００１６】次に配列Ｄと配列Ｅを一つのグループとし、これらと他の各配列との相違度を同様に式（１）により計算する。例えば、配列Ｄ，Ｅと配列Ａとの相違度$V_{(DE)A}$は、$V_{(DE)A}=(V_{AD}+V_{AE})/2$で求められる。同様に、$V_{(DE)B}$，$V_{(DE)C}$についても計算し、これらと、前に求めた$V_{AB}$，$V_{AC}$，・・・から$V_{DE}$を除いたものの中から、最小の値を持つものを選ぶ。この例では、配列Ａと配列Ｂの相違度V_{AB}が最小であり、これらが２番目にグループ化される。以下、同様にグループ化と相違度の計算を行い、その結果をもとに系統樹１３を作成する。

【００１７】図４（Ａ）は、図３に示すアライメントデータをもとに構築された系統樹１３の例であり、図中の括弧内の数字は系統樹１３中の各枝の長さを表している。

（ｃ）各遺伝子配列に対する重み付けの計算
次に、配列の重み計算手段１４は、作成された系統樹１３の各枝の長さをもとに、各枝に重みを付与する。各枝に与える重みは、その枝から分岐した配列の本数で枝の長さを割ることにより求める。

【００１８】図４（Ａ）の系統樹１３を例に説明する。枝１は長さ０．１５８であり、枝１からは配列Ａ、配列Ｂ、配列Ｃの３本の配列が分岐している。従って、枝１の重みは０．１５８／３で０．０５３と求められる。同様にして、全ての枝の重みを求めると、次のようになる。

枝１の重み＝０．１５８／３＝０．０５３
枝２の重み＝０．９０３／２＝０．４５２
枝３の重み＝０．３６７／２＝０．１８４
枝４の重み＝０．７４５
枝５の重み＝枝６の重み＝０．３７８／２＝０．１８９
枝７の重み＝枝８の重み＝０．０００

こうして求めた各枝の重みをもとに各配列の重みを計算する。

【００１９】各配列に付与する重みは、系統樹１３の根（ｒｏｏｔ）から遡った時に通る枝の重み合計として求める。図４（Ａ）の系統樹１３の例では、次のようになる。配列Ａは、系統樹１３上で枝１、枝３、枝５を通る。各枝に与えられた重みは、それぞれ０．０５３，０．１８４，０．３７８である。従って、配列Ａの重みは、これらの合計で０．６１５と求められる。同様にして、全ての配列に対する重みを計算する。更に、全ての配列の重みの合計を求め、その合計で各配列の重みを割り、重みの合計が１になるように標準化する。図４（Ｂ）は、図４（Ａ）の系統樹１３から求めた各配列Ａ～Ｅの重みを示す。

【００２０】（ｄ）各アミノ酸ごとの重み計算
次に、配列の重み計算手段１４は、各配列Ａ～Ｅの標準化された重みをもとに、各部位におけるそれぞれのアミノ酸の重みを求める。そして、部位毎に出現するアミノ酸の重みを、そのアミノ酸が現れるすべての配列の重みの合計として求める。

【００２１】図３に示すアライメントデータと図４（Ｂ）の配列の重みをもとに説明する。第１番目の部位では、配列ＡはＱ（グルタミン）、配列ＢはＬ（ロイシン）、配列ＣはＥ（グルタミン酸）、配列Ｄと配列ＥはＳ（セリン）のアミノ酸がそれぞれ出現している。従って、第１番目の部位では、アミノ酸Ｑの重みには配列Ａの重み０．２１０が与えられ、同様にアミノ酸Ｌには０．２１０、アミノ酸Ｅには０．２７２、アミノ酸Ｓには配列Ｄと配列Ｅの重みの和で０．３０８の重みがそれぞれ与えられる。その他のアミノ酸は、第１番目の部位では重み０となる。

【００２２】同様に、第２番目の部位では、配列Ａ，Ｂ，Ｃにアミノ酸Ｖ（バリン）が出現し、配列Ｄ，Ｅにアミノ酸Ａ（アラニン）が出現している。この部位におけるアミノ酸Ｖの重みは、配列Ａ，Ｂ，Ｃの重みの和で、

０．２１０＋０．２１０＋０．２７２＝０．６９２　となる。第２番目の部位におけるアミノ酸Ａの重みは、０．１５４＋０．１５４＝０．３０８　となる。

【００２３】同様にして、全ての部位において、各アミノ酸の重みを計算する。図５は、以上のようにして計算した第１番目の部位から第１０番目の部位までのアミノ酸の重みを示す。なお、図５では、小数点以下第４桁までの計算結果を示している。図６は、配列の重み計算手段１４が行う、図２に示す処理（ｄ）の一実施例を示すフローチャートである。

【００２４】図６中、ステップ３１は、配列の重み計算手段１４の、即ち、処理装置１０のＣＰＵ内の、部位位置カウンタを初期化する。ステップ３２は、配列の重み計算手段１４の、即ち、処理装置１０のＣＰＵ内の、遺伝子配列番号カウンタを初期化する。ステップ３３は、現部位位置での現遺伝子配列のアミノ酸を、配列の重み計算手段１４の、即ち、処理装置１０のメモリ内の、現アミノ酸種格納領域に格納する。以下の説明で、現部位位置、現遺伝子配列、現アミノ酸種等は、夫々現在注目している部位の位置、現在注目している遺伝子配列、現在注目しているアミノ酸種等を指す。ステップ３４は、現遺伝子配列の重みを、配列の重み計算手段１４の、即ち、処理装置１０のメモリ内の、現部位位置の現アミノ酸種のスコア格納領域に格納されているスコアに加算して格納する。ステップ３５は、遺伝子配列番号カウンタに１を加える。

【００２５】ステップ３６は、遺伝子配列番号カウンタの値が遺伝子数より大きいか否かを判定し、判定結果がＮＯであれば、処理はステップ３４へ戻る。他方、ステップ３６の判定結果がＹＥＳであれば、ステップ３７が現部位位置の各アミノ酸のスコアの合計が１になるように標準化処理を行う。次に、ステップ３８は、各部位でのスコアを計算し、ステップ３９は、部位位置に１を加＊える。ステップ４０は、部位位置カウンタの値がアライメントデータ長より大きいか否かを判定し、判定結果がＹＥＳであれば処理が終了する。他方、ステップ４０の判定結果がＮＯであれば、処理はステップ３２へ戻る。

【００２６】（ｅ）各部位でのスコア計算

配列によっては、性質の類似したアミノ酸への置換が起こっている場合があるが、このような場合でも、機能的に保存されていることが多い。そこで、スコア計算手段１５は、このような部位をモチーフとして抽出するために、アミノ酸間の物理化学的類似性に基づくスコア表をもとに各部位のスコアを計算する。

【００２７】アミノ酸の類似性に基づくスコア表１６ａは、予め各アミノ酸の物理・化学的性質をもとに求められているものであって、各アミノ酸の組の置換頻度や性質の違いの程度を示す距離に基づいて、各アミノ酸の組に対して付与された値を持つテーブルである。例えば、グリシン（Ｇ）と他のアミノ酸との組のスコアは、次のような値が付与されている。ただし、この場合には便宜上各スコアが１００倍されている。

【００２８】

グリシン（Ｇ）	－グリシン（Ｇ）・・・	１００．０
	－アラニン（Ａ）・・・	７４．０
	－セリン（Ｓ）・・・	７５．７
	－ロイシン（Ｌ）・・・	０．０
	・・・・・・	

このようなスコア表については、種々のものが知られているので、ここでの説明はこの程度にとどめる。

【００２９】スコア表１６ａの値を加味して計算した各部位のスコアは、その値が大きいほど、その部位ではアミノ酸が「保存的」であることを示している。例えば、図５の第１番目の部位のスコアは、次の式（２）で求められる。

$$
\begin{aligned}
S_1 = {} & D(S,S)\times S(S)\times S(S) \\
& + D(S,L)\times S(S)\times S(L) \\
& + D(S,E)\times S(S)\times S(E) \\
& + D(S,Q)\times S(S)\times S(Q) \\
& + D(L,S)\times S(L)\times S(S) \\
& + D(L,L)\times S(L)\times S(L) \\
& + D(L,E)\times S(L)\times S(E) \\
& + D(L,Q)\times S(L)\times S(Q) \\
& + D(E,S)\times S(E)\times S(S) \\
& + D(E,L)\times S(E)\times S(L) \\
& + D(E,E)\times S(E)\times S(E) \\
& + D(E,Q)\times S(E)\times S(Q) \\
& + D(Q,S)\times S(Q)\times S(S) \\
& + D(Q,L)\times S(Q)\times S(L) \\
& + D(Q,E)\times S(Q)\times S(E) \\
& + D(Q,Q)\times S(S)\times S(Q) \qquad \cdots \text{式（2）}
\end{aligned}
$$

ここで、S_1 は第一番目の部位のスコア、Ｄ（アミノ酸１，アミノ酸２）はアミノ酸１とアミノ酸２のスコア表

１６ａから得た類似度、Ｓ（アミノ酸）はその部位におけるアミノ酸の重み（図５）である。

【００３０】図７は、スコア計算手段１５が行う、図２に示す処理（ｅ）の一実施例を示すフローチャートである。図７中、ステップ５１は、スコア計算手段１５の、即ち、処理装置１０のメモリ内の、現部位位置のスコア格納領域を初期化する。ステップ５２は、スコア計算手段１５の、即ち、処理装置１０のメモリ内の、アミノ酸種カウンタを初期化する。ステップ５３は、スコア計算手段１５の、即ち、処理装置１０のメモリ内の、比較アミノ酸種カウンタを初期化する。ステップ５４は、現アミノ酸種と現比較アミノ酸種間の類似度を、アミノ酸類似度スコア表１６ａを参照して得る。ステップ５５は、

$S_i = S_i + D(A_1, A_2) \times S(A_1) \times S(A_2)$

なる計算を行う。ここで、S_i は現部位位置（ｉ番目）のスコア、A_1 は現アミノ酸種、A_2 は現比較アミノ酸種、$D(A_1, A_2)$ は現アミノ酸種と現比較アミノ酸＊種の類似度のスコア、$S(A_1)$ は現部位位置の現アミノ酸種のスコア、$S(A_2)$ は現部位位置の現比較アミノ酸種のスコアを夫々示す。

【００３１】ステップ５６は、比較アミノ酸種を次の比較アミノ酸種に変更し、ステップ５７は、全てのアミノ酸種との比較が行われたか否かを判定する。ステップ５７の判定結果がＮＯであれば、処理はステップ５３へ戻る。他方、ステップ５７の判定結果がＹＥＳであると、ステップ５８でアミノ酸種を次のアミノ酸種に変更する。ステップ５９は、全てのアミノ酸種についてスコアの計算を行ったか否かを判定し、判定結果がＹＥＳであれば処理が終了する。他方、ステップ５９の判定結果がＮＯであれば、処理はステップ５２へ戻る。

【００３２】（ｆ）計算結果出力

図３に示すアライメントデータについて、スコア計算手段１５でスコアを計算した結果は以下のとおりであった。

（部位０１～０５）	0.5183	0.7744	0.5677	0.8198	0.4881
（部位０６～１０）	0.9328	0.4940	0.8683	0.3165	0.3580
（部位１１～１５）	0.9311	0.3834	0.4072	0.3611	0.6114
（部位１６～２０）	0.6937	0.5976	0.5699	0.5574	0.5010
（部位２１～２５）	0.3880	0.6168	0.5530	0.5739	0.6296
（部位２６～３０）	0.7718	0.3473	0.3772	0.6956	1.0000
（部位３１～３５）	0.9841	0.9646	1.0000	0.9149	0.8891
（部位３６～４０）	1.0000	0.6916	0.7864	0.7804	0.7903
（部位４１～４５）	0.5830	0.6021	0.7753	0.5654	0.6976
（部位４６～５０）	0.9037	0.6428	0.8303	0.9542	0.7105

（ｇ）閾値設定

特徴情報抽出手段１７は、スコアに閾値を決定し、その閾値を超えるスコアの与えられた部位をモチーフとして抽出する。そのため、閾値を、ユーザの指定またはディフォルト値として事前に定められている値により設定する。

【００３３】（ｈ）閾値を超える部位をモチーフとして出力

スコアの閾値がｔｈの場合、特徴情報抽出手段１７は、次式の条件を満たす部位をモチーフの候補として抽出する。

$S > th$

図３に示すアライメントデータについて、スコアの閾値ｔｈを０．９０として抽出したモチーフは、以下のとおりであった。

【００３４】『３０　Ｄ［ＬＩ］［ＩＭ］Ｌ［ＬＩＦ］［ＫＲＨ］Ｌ』

ここで、「３０」は図３のアライメントデータ中におけるモチーフの先頭アミノ酸の位置が３０であることを意味する。また、［］は、その部位では［］内の複数のアミノ酸が出現していることを示す。即ち、抽出されたモチーフ部位は、アライメントデータ中の６番目の（ＦまたはＹ）、１１番目（ＦまたはＹ）、３０番目（Ｄ）、３１番目（ＬまたはＩ）、３２番目（ＩまたはＭ）、３３番目（Ｌ）、３４番目（ＬまたはＩまたはＦ）、３６番目（Ｌ）、４６番目（ＩまたはＶ）及び４９番目（ＬまたはＩまたはＭ）の部位である。

【００３５】図８は、本発明の実施例によって抽出されたモチーフ部位を示す図である。ところで、従来は、遺伝子配列に特徴的な配列パターンであるモチーフを部位としてマニュアル操作である程度までは抽出することができたが、モチーフを領域として同定することは困難であった。しかし、機能領域の同定や、祖先遺伝子の推定を行う場合、モチーフを領域として同定することは非常に重要である。そこで、本発明において、部位として抽出されたモチーフ配列を領域として同定する方法について、より詳細に説明する。

【００３６】（ｉ），（ｊ），（ｋ）領域幅及びランダムレベルの設定、モチーフ部位の出現率の計算、モチーフ領域の出力

図９は、特徴情報抽出手段１７が行う、図２に示す処理（ｉ），（ｊ），（ｋ）の一実施例を示すフローチャートである。本実施例は、大略３つの処理からなる。第１の処理では、任意の領域幅を設定し、その領域幅内のモ

チーフ部位の出現率を求める。第２の処理では、設定した領域幅内でのモチーフ部位の出現率が充分に高いか否かを判断するためのランダムレベルを求め、ランダムレベルを越える出現率でモチーフ部位が存在する場合はその領域幅内のモチーフ部位を１つのモチーフ領域として同定する。第３の処理では、同定されたモチーフ領域が連続する場合にはそれらをまとめて１つのモチーフ領域とする。

【００３７】つまり、より具体的には以下の処理Ｓ１～Ｓ６が繰り返される。

Ｓ１：モチーフ部位の抽出を行う。

Ｓ２：初期領域幅、拡張幅、最大拡張幅を設定する。また、モチーフ部位の出現率のランダムレベルを求めるための領域幅を最大拡張幅に設定する。ただし、最大拡張幅がアライメントデータ長の半分を越える場合には、ランダムレベルの領域幅をアライメントデータ長の半分の長さを越えない値に設定する。

Ｓ３：初期領域幅及びランダムレベルの領域幅の夫々でのモチーフ部位の出現率を計算してプロットする。

Ｓ４：初期領域幅でのモチーフ部位の出現率がランダムレベルの領域幅でのモチーフ部位の出現率を越えている場合には、初期領域幅を「モチーフ領域」とみなす。

Ｓ５：隣合う初期領域幅のモチーフ部位の出現率がともに「モチーフ領域」である場合には、これらを結合して１つの「モチーフ領域」とみなす。

Ｓ６：処理Ｓ４及びＳ５をアライメントデータの全長に渡って繰り返す。

【００３８】図９に基づいてモチーフ領域の同定処理を説明するに、ステップ６１はモチーフ部位の出現率を求める領域幅を設定する。ステップ６２は、特徴情報抽出手段１７の、即ち、処理装置１０のＣＰＵ内の、部位位置カウンタを初期化する。ステップ６３は、現部位位置を中心として、設定した領域幅内でのモチーフ部位の出現率を、次の式から求める。

【００３９】（モチーフ部位の出現率）＝（領域幅内モチーフ部位数）／（領域幅）

ステップ６４は、モチーフ部位の出現率をグラフにプロットし、ステップ６５は、現部位位置でのランダムレベルを計算し、グラフにプロットする。ステップ６６は、部位位置に１を加え、ステップ６７は、部位位置カウンタの値がアライメントデータ長より大きいか否かを判定する。ステップ６７の判定結果がＮＯであれば、処理はステップ６３へ戻る。

【００４０】他方、ステップ６７の判定結果がＹＥＳであると、ステップ６８で特徴情報抽出手段１７の、即ち、処理装置１０のＣＰＵ内の、モチーフ領域フラグを初期化する。ステップ６９は、部位位置カウンタを初期化する。ステップ７０は、部位位置のモチーフ出現率がランダムレベルより高いか否かを判定する。ステップ７０の判定結果がＹＥＳであれば処理はステップ７１へ進み、ＮＯであれば処理はステップ７５へ進む。

【００４１】ステップ７１は、モチーフ領域フラグが立っているか（セットされているか）否かを判定し、判定結果がＹＥＳであると、ステップ７２が現部位位置を中心とした領域を現モチーフ領域に加えて伸長する。他方、ステップＳ７１の判定結果がＮＯであると、ステップ７３はモチーフ領域フラグを立て、ステップ７４は、現部位位置を中心とした領域幅の中で最初にモチーフ部位の出現する部位位置を現モチーフ領域の開始部位とする。ステップ７２又は７４を行った後は、処理がステップ７８へ進む。

【００４２】ステップ７５は、モチーフ領域フラグが立っているか否かを判定し、判定結果がＮＯであれば、処理はステップ７８へ進む。他方、ステップ７５の判定結果がＹＥＳであれば、ステップ７６がモチーフ領域フラグを初期化し、スッテプ７７が現モチーフ領域を出力する。このステップ７６を行った後は、処理がステップ７８へ進む。

【００４３】ステップ７８は、部位位置に１を加える。又、ステップ７９は、部位位置カウンタの値がアライメントデータ長より大きいか否かを判定し、判定結果がＹＥＳであれば、処理は終了する。他方、ステップ７９の判定結果がＮＯであれば、処理はステップ７０へ戻る。上記の如きモチーフ領域の同定処理を行った場合の実験結果を以下に説明する。

【００４４】実験では、ＦＬＡＡ７Ａ－１をプローブとして、アライメントデータを対象としてモチーフ領域の同定を行った。図１０～図１２は、プローブ名がＦＬＡＡ７Ａ－１、ｈｏｍｏｌｏｇｕｅ本数が５３、初期領域幅が２１、最大拡張幅が１０１、アライメントデータ長が９７、ランダムレベルを求めるための領域幅が４１、モチーフ部位抽出時の設定値が０．９０の場合の実験結果を示す。図１０は、設定した領域幅におけるモチーフ部位の占める割合を示しており、「ｏ」はモチーフ領域幅の初期値でのプロットを示し、「．．．」はランダムレベルのプロットを示す。尚、プロットが重なった場合は、割合の高い方を優先してプロットしてある。図１１は、抽出されたモチーフ部位を示す。更に、図１２は、モチーフ領域の同定処理により得られたモチーフ領域を示している。図１２中、「：」はモチーフの開始位置と終了位置とを示し、「［　］」はそのモチーフ部位に複数のアミノ酸が出現することを示し、「－」はその部位に任意のアミノ酸又はギャップ（即ち、モチーフ部位ではない部位）が出現することを示す。

【００４５】図１３～図１５は、同様にしてＥＣＯＤＨＦＯＬＧ－１をプローブとして用いて得られた他の実験結果を示す。図１３～図１５は、プローブ名がＥＣＯＤＨＦＯＬＧ－１、ｈｏｍｏｌｏｇｕｅ、本数が１１、初期領域幅が１１、アライメントデータ長が１７９、ランダムレベルを求めるための領域幅が８１、モチーフ部位

抽出時の設定値が０．９０の場合の実験結果を示す。図１３及び図１４は、同じアライメントデータに対するプロットを分割して示しており、図１５は、モチーフ領域の同定処理により得られたモチーフ領域を示す。 図１３及び図１４は、設定した領域幅におけるモチーフ部位の占める割合をアライメントデータと対応させて示しており、「o」はモチーフ領域幅の初期値でのプロット、即ち、設定した領域幅でのモチーフ部位の出現率を示し、「．．．」はランダムレベルのプロットを示す。つまり、モチーフ部位の出現率がこのランダムレベルより低い場合は、このモチーフ部位をモチーフ領域とはみなさない。尚、プロットが重なった場合は、割合の高い方を優先してプロットしてある。更に、図１３及び図１４中、アライメントデータの左側に示されている名前は、遺伝子配列データベースＤＤＢＪに登録されている遺伝子配列のエントリー名を示す。又、Ｄｉｈｙｄｒｏｆｏｌａｔｅ ｒｅｄｕｃｔａｓｅ ｓｉｇｎａｔｕｒｅ［ＬＩＦ］－Ｇ－Ｘ（４）－［ＬＩＶＭＦ］－Ｐ－Ｗは、モチーフデータベースＰＲＯＳＩＴＥに登録されているデータである。

【００４６】図１５中、左側に示されている「１２２」等の数字は、各モチーフ領域のアライメントデータ上での開始位置を示す。又、右側に示されている「１３７」等の数字は、各モチーフ領域のアライメントデータ上での終了位置を示す。図１６～図１８は、同様にしてＨＵＭＴＲＸ１－１をプローブとして用いて得られた他の実験結果を示す。図１６～図１８は、プローブ名がＨＵＭＴＲＸ１－１、ｈｏｍｏｌｏｇｕｅ、本数が１５、初期領域幅が１１、アライメントデータ長が１１０、ランダムレベルを求めるための領域幅が５１、モチーフ部位抽出時の設定値が０．９０の場合の実験結果を示す。図１６及び図１７は、同じアライメントデータに対するプロットを分割して示しており、図１８は、モチーフ領域の同定処理により得られたモチーフ領域を示す。

【００４７】図１６及び図１７は、設定した領域幅におけるモチーフ部位の占める割合をアライメントデータと対応させて示しており、「o」はモチーフ領域幅の初期値でのプロット、即ち、設定した領域幅でのモチーフ部位の出現率を示し、「．．．」はランダムレベルのプロットを示す。つまり、モチーフ部位の出現率がこのランダムレベルより低い場合は、このモチーフ部位をモチーフ領域とはみなさない。尚、プロットが重なった場合は、割合の高い方を優先してプロットしてある。更に、図１６及び図１７中、アライメントデータの左側に示されている名前は、遺伝子配列データベースＤＤＢＪに登録されている遺伝子配列のエントリー名を示す。又、Ｔｈｉｏｒｅｄｏｘｉｎ ｆａｍｉｌｙ ａｃｔｉｖｅ ｓｉｔｅ［ＳＴＡ］－Ｘ－［ＷＧ］－Ｃ－［ＡＧＶ］－［ＰＨ］－Ｃは、モチーフデータベースＰＲＯＳＩＴＥに登録されているデータである。

【００４８】図１８中、左側に示されている「６９」等の数字は、各モチーフ領域のアライメントデータ上での開始位置を示す。又、右側に示されている「１０５」等の数字は、各モチーフ領域のアライメントデータ上での終了位置を示す。これにより、本発明によれば、遺伝子配列情報から、機械的に（自動的に）モチーフ領域を抽出・同定することができるので、高速にモチーフ領域の抽出・同定が可能となる。従って、大量の遺伝子配列データから新規なモチーフを発見したり、モチーフデータベースを作成することが、容易にできる。この様にして得られたモチーフ情報をもとに、未知機能の遺伝子配列の機能及び構造の予測を効率良く行うことができ、本発明を遺伝子機能の発見や機能領域の同定に利用すると非常に便利である。

【００４９】以上、本発明を実施例により説明したが、本発明は上記実施例に限定されるものではなく、本発明の範囲内で種々の変形及び改良が可能であることは、言うまでもない。

【００５０】

【発明の効果】以上説明したように、本発明によれば、遺伝子配列情報から、機械的に（自動的に）モチーフ領域を抽出・同定することができるので、高速にモチーフ領域の抽出・同定が可能となるので、大量の遺伝子配列データから新規なモチーフを発見したり、モチーフデータベースを作成することが、容易にでき、この様にして得られたモチーフ情報をもとに、未知機能の遺伝子配列の機能及び構造の予測を効率良く行うこともできるので、本発明を遺伝子機能の発見や機能領域の同定に利用すると非常に便利であり、遺伝子工学の発展に寄与するところが大きい。

【図面の簡単な説明】

【図１】本発明の構成例を示す図である。

【図２】本発明の実施例の処理フローチャートである。

【図３】入力したアライメントデータの例を示す図である。

【図４】図３に示すアライメントデータから求めた系統樹と配列の重みの結果を示す図である。

【図５】各部位におけるアミノ酸の重みの計算結果を示す図である。

【図６】配列の重み計算手段が行う処理の一実施例を示すフローチャートである。

【図７】スコア計算手段が行う処理の一実施例を示すフローチャートである。

【図８】本発明の実施例によって抽出されたモチーフ部位を示す図である。

【図９】特徴情報抽出手段が行う処理の一実施例を示すフローチャートである。

【図１０】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その１）。

【図１１】モチーフ領域の同定処理を行った場合の実験

結果を説明する図である（その２）。

【図１２】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その３）。

【図１３】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その１）。

【図１４】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その２）。

【図１５】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その３）。

【図１６】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その１）。

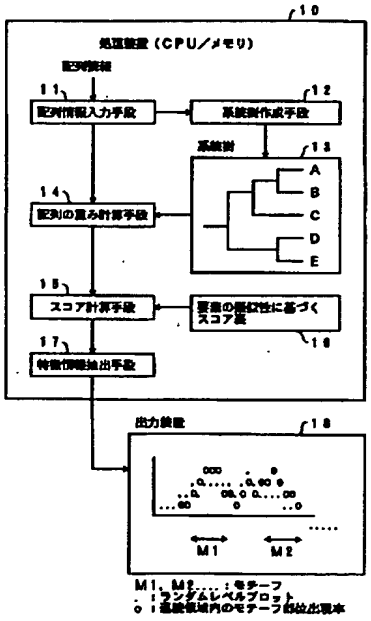
【図１７】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その２）。

【図１８】モチーフ領域の同定処理を行った場合の実験結果を説明する図である（その３）。

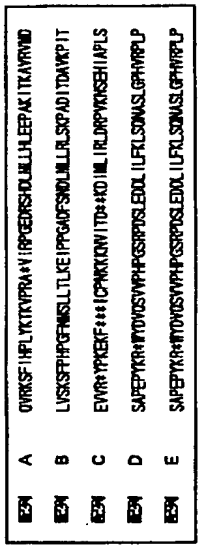
【符号の説明】

１０　処理装置
１１　配列情報入力手段
１２　系統樹作成手段
１３　系統樹
１４　配列の重み計算手段
１５　スコア計算手段
１６　要素の類似性に基づくスコア表
１７　特徴情報抽出手段
１８　出力装置

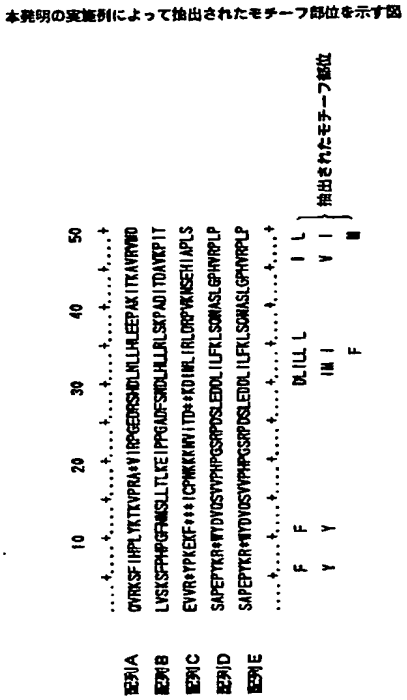
【図１】



【図３】



【図８】



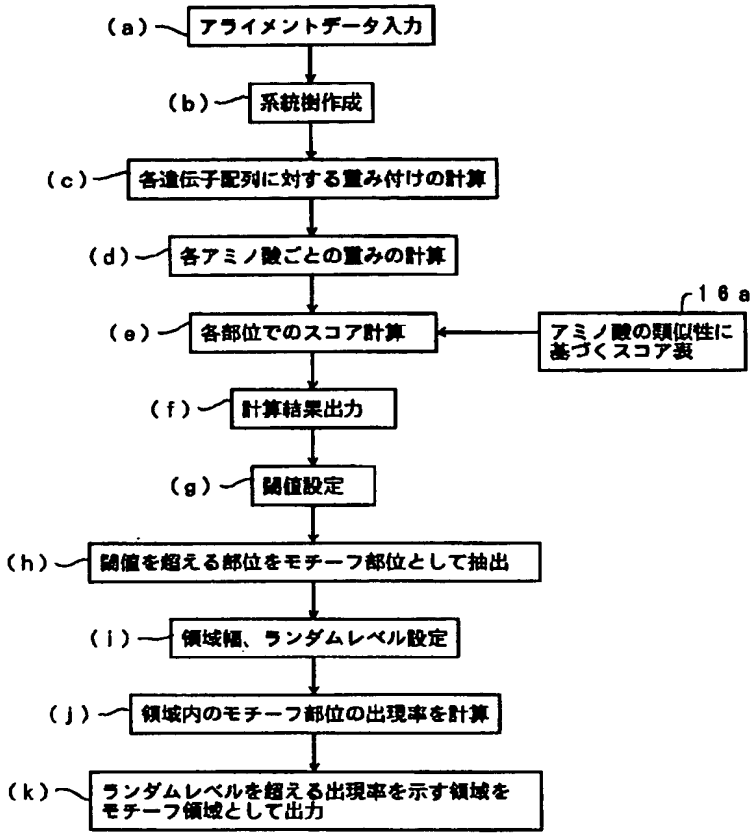
【図１８】

モチーフ領域の同定処理を行った場合の実験結果を
説明する図（その３）

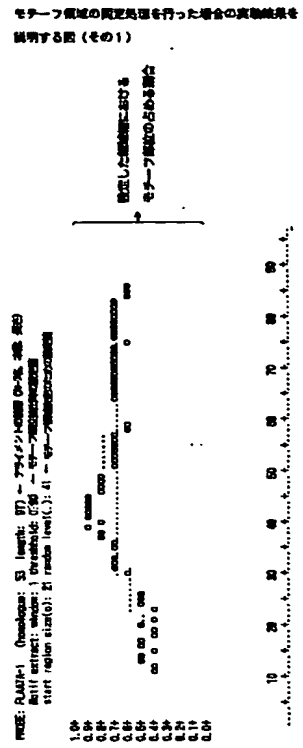
```
13: [FY] :13
25: [IV]--DF-A-RCGPC[KR]-[IV]-P----[LIF] :47
61: [DN]-D-----[AP] :89
69: [AP]-------[TIN]P[AT][LF]--[LFY]K-G-------G------[LI]---[LIV] :105
```

【図2】

本発明の実施例の処理フローチャート

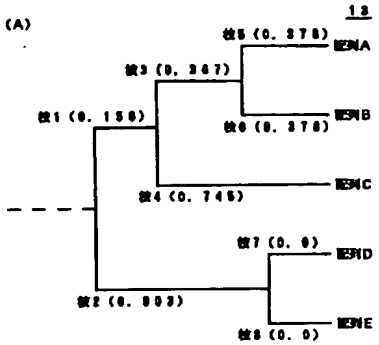


【図10】



【図4】

図3に示すアライメントデータから求めた系統樹と配列の重みの結果を示す図



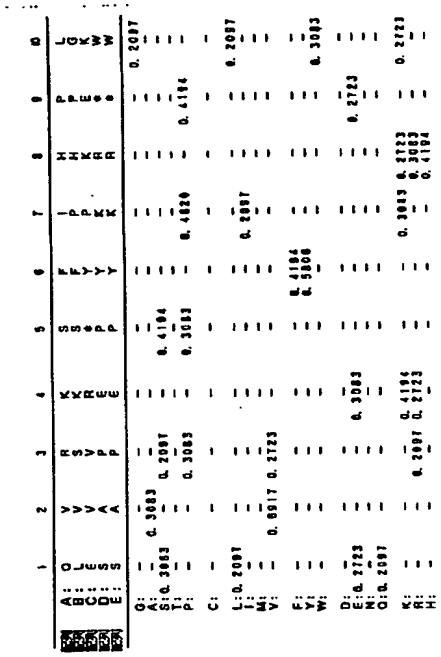
( ) 内の数字は枝の長さを示す。

(B) 各配列の重み

		標準化
配列A	0. 815	0. 210
配列B	0. 815	0. 210
配列C	0. 796	0. 272
配列D	0. 452	0. 154
配列E	0. 452	0. 154

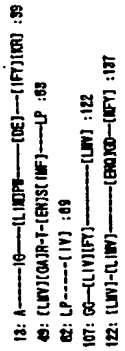
【図5】

各部位におけるアミノ酸の重みの計算結果を示す図



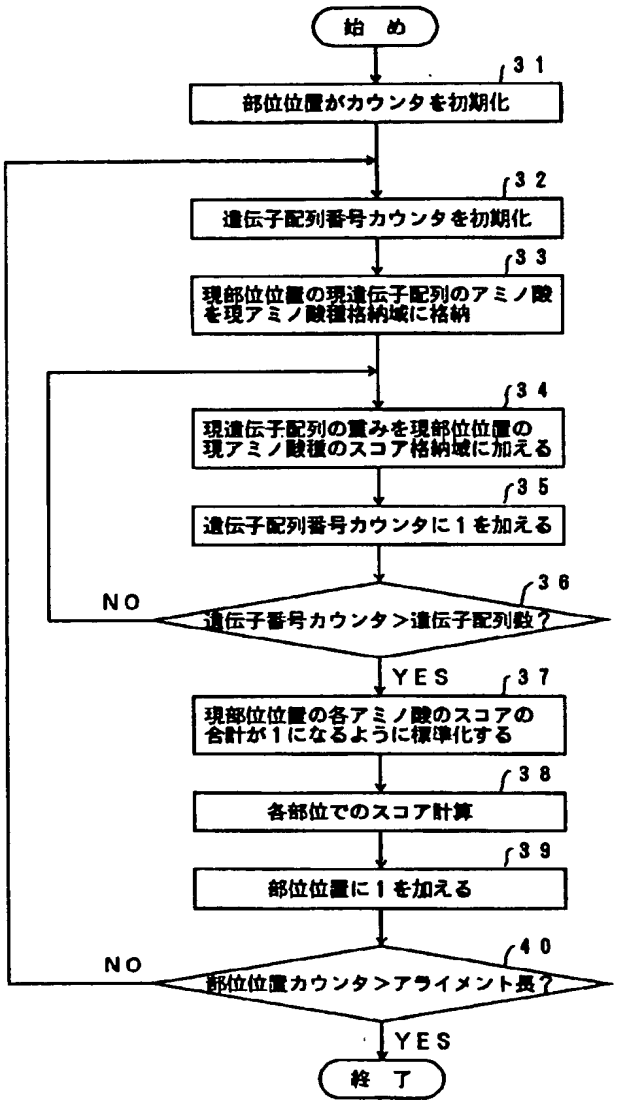
【図15】

モチーフ領域の同定処理を行った場合の実験結果を説明する図（その3）



【図6】

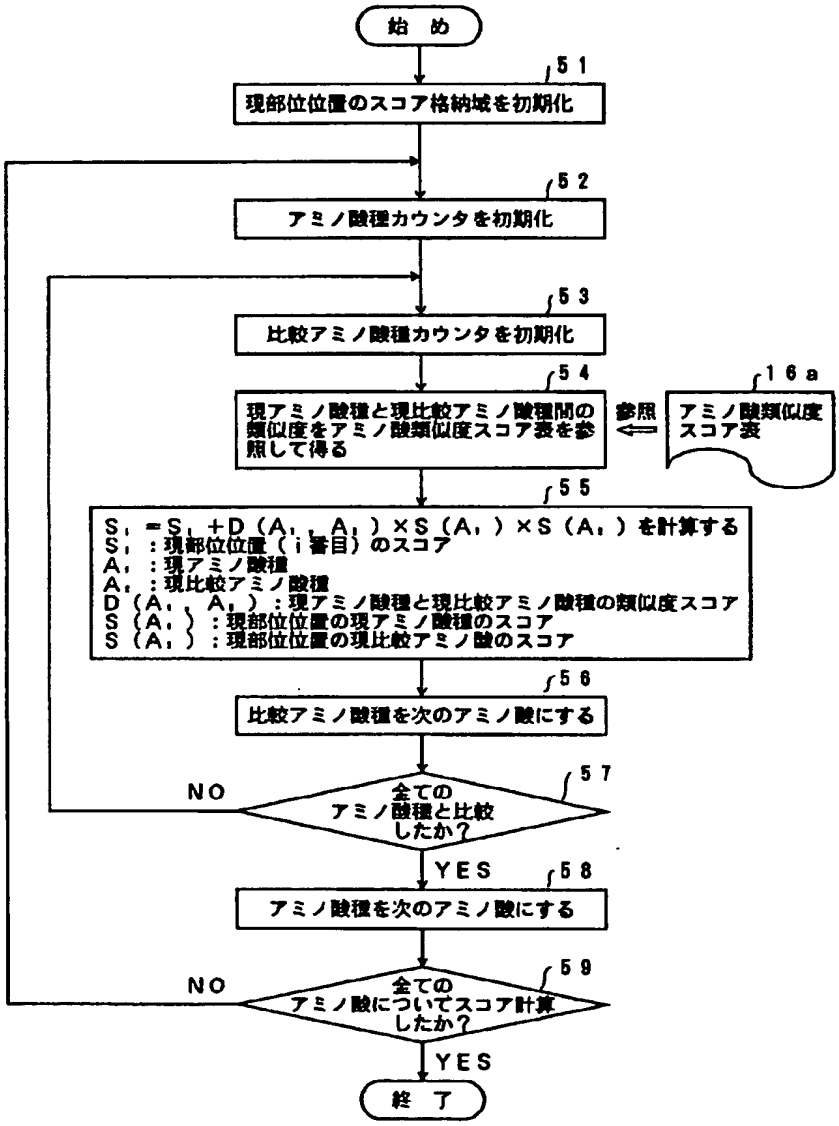
配列の置み計算手段が行う処理の一実施例を示すフローチャート



【図１１】

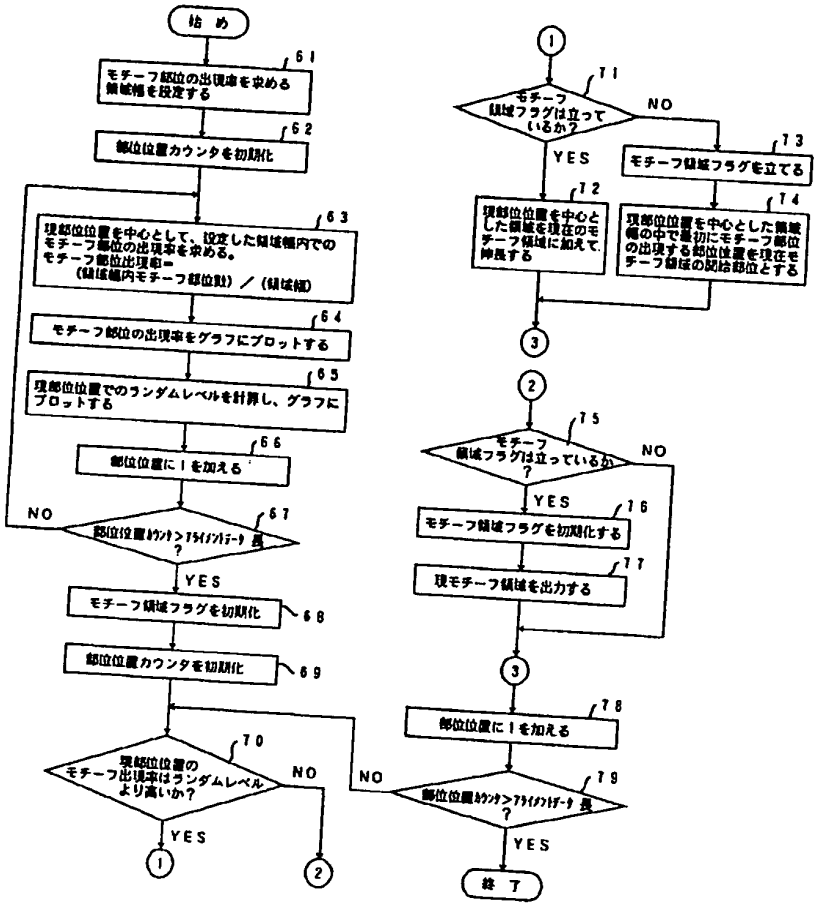
【図7】

スコア計算手段が行う処理の一実施例を示すフローチャート



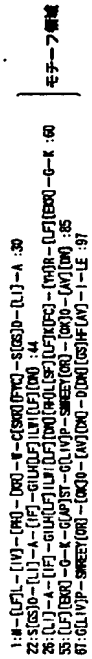
【図9】

特徴情報抽出手段が行う処理の一実施例を示すフローチャート



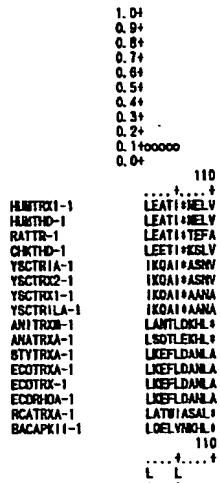
【図１２】

モチーフ領域の同定処理を行った場合の実験結果を説明する図（その３）



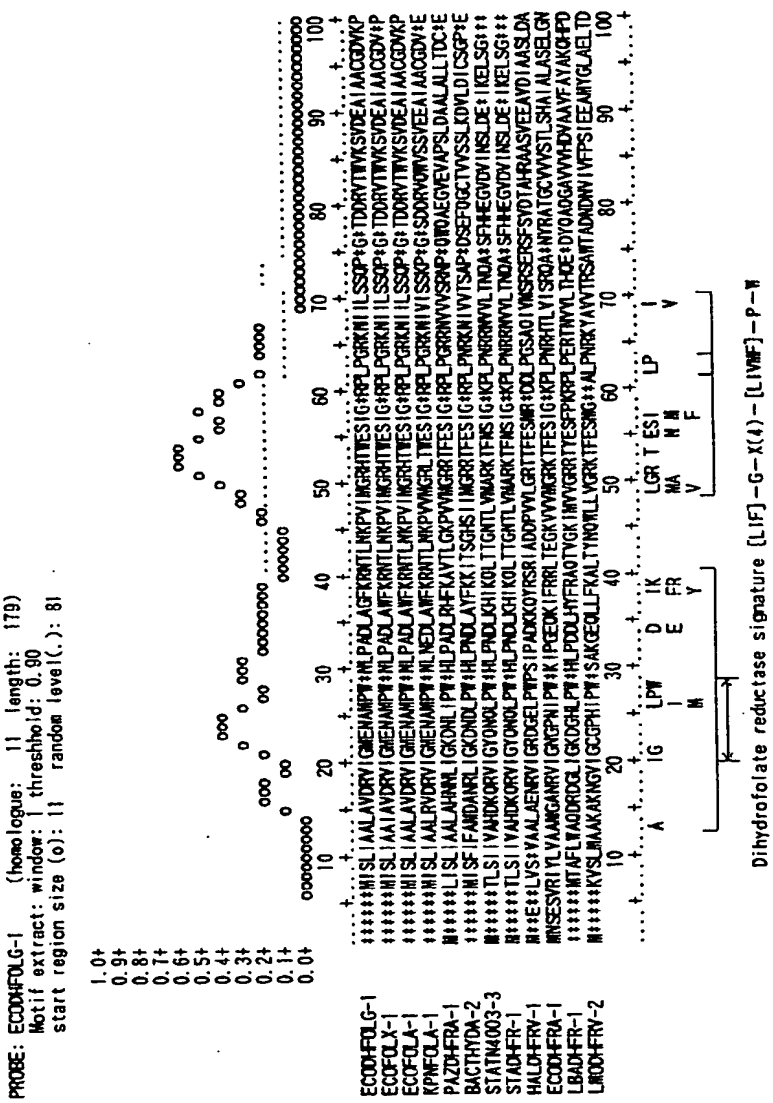
【図１７】

モチーフ領域の同定処理を行った場合の実験結果を説明する図（その２）



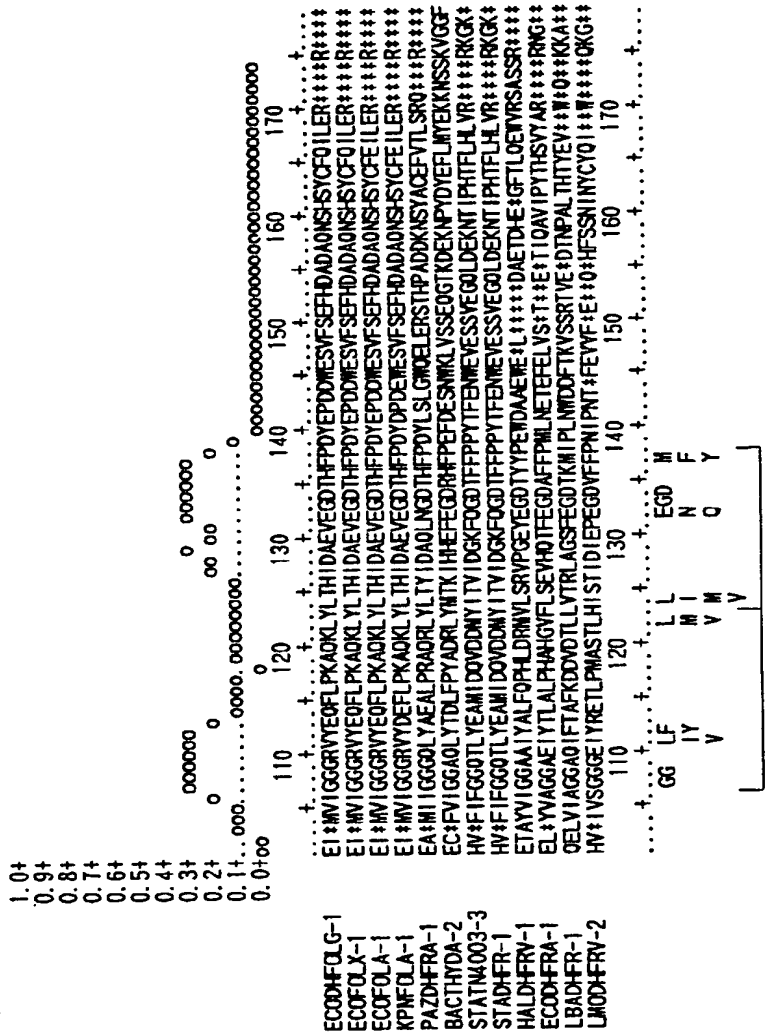
【図１３】

モチーフ領域の同定処理を行った場合の実験結果を説明する図（その１）



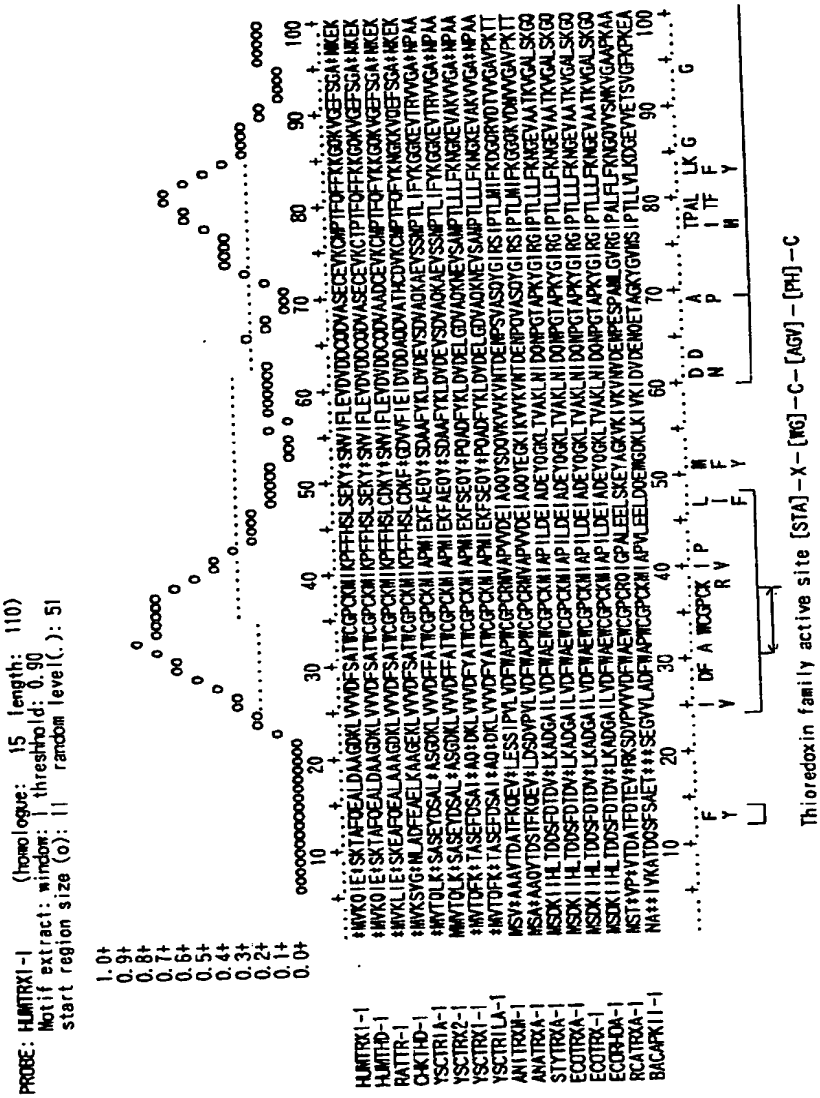
【図１４】

モチーフ領域の同定処理を行った場合の実験結果を
説明する図（その２）



【図１６】

モチーフ領域の同定処理を行った場合の実験結果を説明する図（その１）



フロントページの続き

(72)発明者　池尾　一穂
静岡県三島市谷田1111番地　国立遺伝学研究所内

(72)発明者　川西　祐一
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内

(72)発明者　河合　正人
神奈川県川崎市中原区上小田中1015番地
富士通株式会社内